FINAL PROJECT: EXPLORASI DAN VISUALISASI DATA

Student ID	Student name	Contribution description	Contribution (%)
2106651572	Rafly Witjaksana	e.g coding, perumusan masalah, penulisan laporan	100%
2106635745	Muhamad Rakan Akmal	coding, aktif dalam diskusi kelompok, pre-processing data.	100%
2106701974	Rima Fitrianti Azahra	Aktif dalam diskusi kelompok, data gathering, perumusan masalah.	100%
2106707864	Tsabita Asir Saladin	Aktif dalam diskusi kelompok, Membuat materi presentasi	100%

LINK FOLDER DRIVE:
https://drive.google.com/drive/folders/1jKghNEFKoswjYjPcuA9HQSYhZQdvEhxL?usp=sharing

PENDAHULUAN

Berdasarkan data survei yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) mengenai pengguna internet di Indonesia, sebanyak 69,64% menggunakan internet untuk melihat video dan 70,23% menggunakan internet untuk mengunduh video dari seluruh pengguna internet di Indonesia.

Beberapa tahun terakhir, layanan SVOD (Subscription Video on Demand) banyak bermunculan. Layanan SVOD adalah layanan TV atau film berlangganan yang dapat dinikmati oleh pelanggan dengan menggunakan jaringan internet. Beberapa diantaranya adalah Netflix, Hulu, Prime Video, dan Disney+. Untuk dapat berlangganan, tentunya layanan SVOD menetapkan biaya untuk mengakses film atau layanan TV yang tidak dapat dikatakan murah. Sehingga memunculkan permasalahan dalam memilih SVOD untuk seseorang, agar layanan yang didapatkan cocok dengan hal yang mereka suka.

Untuk menentukan SVOD mana yang paling cocok untuk berbagai kalangan orang, kami akan melakukan analisis berdasarkan jumlah film, tahun rilis film, batasan umur penonton film, rating IMDb film, genre film, negara asal film, bahasa film dari suatu SVOD serta trend analysis berdasarkan data text.

Data CSV yang digunakan pada analisis ini, diambil dari:
https://raw.githubusercontent.com/smartinternz02/SBSPS-Challenge-5372-OTT-Platform-Analysis-Tool/main/MoviesOnStreamingPlatforms_updated.csv

Ukuran Data: 16744 row x 16 column

Didapat bahwa data berisi Variabel berikut:

- ID = Kode Unik Film (type: int64)
- Title = Judul Film (type: object)
- Year = Tahun Rilis Film (type: int64)

- Age = Batasan Usia Penonton Film (type: object)
- IMDb = Rating IMDb Film (type: float64)
- Rotten Tomatoes = Rating Rotten Tomatoes Film (type: object)
- Netflix = Apakah film itu ada di Netflix (type: object)
- Hulu = Apakah film itu ada di Hulu (type: object)
- Prime Video = Apakah film itu ada di Prime Video (type: object)
- Disney+ = Apakah film itu ada di Disney+ (type: object)
- Type = Film atau Serial TV (type: int64)
- Directors = Sutradara Film (type: object)
- Genres = Genre Film (type: object)
- Country = Asal Negara Film (type: object)
- Language = Bahasa Film (type: object)
- Runtime = Durasi Film (type: float64)

Gambaran Data:

	ID	Title	Year	Age	IMDb	Rotten Tomatoes	Netflix	Hulu	Prime Video	Disney+	Type	Directors	Genres	Country	Language	Runtime
0	1	Inception	2010	13+	8.8	87%	On Netflix	Not on Hulu	Not on Amazon Prime	Not onn Diney+	0	Christopher Nolan	Action,Adventure,Sci-Fi,Thriller	United States,United Kingdom	English,Japanese,French	148.0
1	2	The Matrix	1999	18+	8.7	87%	On Netflix	Not on Hulu	Not on Amazon Prime	Not onn Diney+	0	Lana Wachowski,Lilly Wachowski	Action,Sci-Fi	United States	English	136.0
2	3	Avengers: Infinity War	2018	13+	8.5	84%	On Netflix	Not on Hulu	Not on Amazon Prime	Not onn Diney+	0	Anthony Russo,Joe Russo	Action,Adventure,Sci-Fi	United States	English	149.0
3	4	Back to the Future	1985	7+	8.5	96%	On Netflix	Not on Hulu	Not on Amazon Prime	Not onn Diney+	0	Robert Zemeckis	Adventure,Comedy,Sci-Fi	United States	English	116.0
4	5	The Good, the Bad and the Ugly	1966	18+	8.8	97%	On Netflix	Not on Hulu	On Amazon Prime	Not onn Diney+	0	Sergio Leone	Western	Italy,Spain,West Germany	Italian	161.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16739	16740	The Ghosts of Buxley Hall	1980	NaN	6.2	NaN	Not on Netflix	Not on Hulu	Not on Amazon Prime	On Diney+	0	Bruce Bilson	Comedy,Family,Fantasy,Horror	United States	English	120.0
16740	16741	The Poof Point	2001	7+	4.7	NaN	Not on Netflix	Not on Hulu	Not on Amazon Prime	On Diney+	0	Neal Israel	Comedy,Family,Sci-Fi	United States	English	90.0
16741	16742	Sharks of Lost Island	2013	NaN	5.7	NaN	Not on Netflix	Not on Hulu	Not on Amazon Prime	On Diney+	0	Neil Gelinas	Documentary	United States	English	NaN
16742	16743	Man Among Cheetahs	2017	NaN	6.6	NaN	Not on Netflix	Not on Hulu	Not on Amazon Prime	On Diney+	0	Richard Slater-Jones	Documentary	United States	English	NaN
16743	16744	In Beaver Valley	1950	NaN	NaN	NaN	Not on Netflix	Not on Hulu	Not on Amazon Prime	On Diney+	0	James Algar	Documentary,Short,Family	United States	English	32.0

16744 rows × 16 columns

Selain itu, data text diambil dari API Twitter dengan gambaran data sebagai berikut:

```
tweets[:7]
```

```
['Kemarin Netflix baru saja merilis Teaser untuk lanjutan episode season 1 dari The Cuphead Show!. Episode baru bakal rilis pada 19 Agustus 2022.\n#TheCuph
 'RT @piargh: NO SMART TV? NO PROBLEM!\n\nasalkan ada wifi and tv, korang boleh layan netflix, disney+ hotstar, prime video, youtube &amp;amp; semua la…',
 'RT @piargh: NO SMART TV? NO PROBLEM!\n\nasalkan ada wifi and tv, korang boleh layan netflix, disney+ hotstar, prime video, youtube &amp;amp; semua la…',
 'RT @lovelyagvrl: HELP RT !\nhalo aku open order apps premium buat ongkos ke rs minggu ini. Aku open order netflix, disney, viu, wetv, iqiyi,…',
 '@ngelaugh Netflix 1p1u 33k/bulan kakk\nvideo shar 16k/bulan \nyuk kak',
 'RT @lovelyagvrl: HELP RT !\nhalo aku open order apps premium buat ongkos ke rs minggu ini. Aku open order netflix, disney, viu, wetv, iqiyi,…',
 'RT @lovelyagvrl: HELP RT !\nhalo aku open order apps premium buat ongkos ke rs minggu ini. Aku open order netflix, disney, viu, wetv, iqiyi,…']
```

PRE-PROCESSING

Data CSV yang kami gunakan untuk melakukan analisis, sudah memiliki persebaran yang dapat dikatakan normal melalui data.describe(). Setelah melihat column SVOD, yaitu Netflix, Hulu, Prime Video, dan Disney+ yang menjadi fokus utama analisis, terdapat permasalahan yang dapat menyulitkan analisis yang kami akan lakukan. Adanya perbedaan terkait values yang menunjukkan bahwa film terdapat pada salah satu SVOD, seperti ‘On Netflix’ yaitu film yang terdapat di SVOD Netflix, ‘On Hulu’ yang menunjukkan film terdapat di SVOD Hulu, akan menyulitkan dalam perhitungan jumlah karena tidak berbentuk angka, dan penyamarataan suatu kondisi. Untuk memudahkan analisis terkait masing-masing SVOD, kami memutuskan untuk melakukan replace values yang menunjukkan bahwa film terdapat

pada salah satu SVOD dengan nilai 1 atau 0 (1 menunjukkan film terdapat pada SVOD kolom tersebut dan 0 menunjukkan film tidak terdapat pada SVOD kolom tersebut). Hal ini dilakukan agar kondisi bahwa film terdapat atau tidaknya pada suatu SVOD dinyatakan dengan nilai yang sama, yaitu 1 atau 0 dan perhitungan jumlah untuk visualisasi dapat dilakukan jauh lebih mudah karena values berupa suatu angka.

Kami melakukan cek missing values pada keseluruhan data dan melihat persebaran missing values dengan menggunakan heatmap. Didapat bahwa colum "Rotten Tomatoes" memiliki jumlah missing value yang paling banyak diantara variabel lainnya. Dengan pertimbangan nilai rating yang digunakan masih dapat didasarkan pada variabel "IMDb", variabel "Rotten Tomatoes" akan di drop. Missing values juga terdapat pada column ‘Age’, ‘IMDb, ‘Directors’, ‘Genres’, ‘Country’, ‘Language’, dan ‘Runtime’. Tetapi karena variabel tersebut merupakan suatu jenis variabel yang tidak dapat dilakukan imputasi dengan value tertentu dan juga tidak dapat dilakukan asumsi yang berkaitan dengan variabel lainnya, kami memutuskan untuk membiarkan missing values pada column-column tersebut.

Setelah melihat secara lebih dalam mengenai column-column pada data, kami menemukan bahwa terdapat satu column, yaitu column ‘Type’, hanya memiliki 1 values, yaitu 0. Column ‘Type’ seharusnya menunjukkan bahwa judul tersebut merupakan film atau serial, namun terlihat bahwa semuanya berjenis '0' yang tidak menggambarkan apapun. Maka, kolom ini akan di drop dengan pertimbangan bahwa kolom ini tidak dapat digunakan untuk visualisasi.

Sedangkan, pada **DATA TEXT** yang kami dapat dari API Keys Twitter, kami melakukan preprocessing data cleaning menggunakan stopwords didasarkan pada bahasa indonesia dengan module taudata. Setelah itu, kami melakukan filtering kata-kata yang kami butuhkan untuk visualisasi nantinya yang berkaitan dengan SVOD yang kami analisa (Netflix, Disney+, Prime Video, dan Hulu).

PENGOLAHAN DAN ANALISIS DATA

Sebelum melakukan analisis data yang lebih jauh, kami menjumlahkan data untuk setiap SVOD yang akan dilakukan analisis. Dari informasi tersebut terlihat bahwa prime video memiliki jumlah film paling banyak di antara platform lainnya. Untuk melihat perbandingan hasil data tersebut, kami melakukan plotting dengan menggunakan barplot.

Setelah melihat perbandingan data secara jumlah dari setiap SVOD, kami mulai melakukan analisis yang berkaitan dengan kelebihan-kelebihan yang dimiliki setiap SVOD untuk membantu menentukan SVOD mana yang cocok untuk suatu kalangan orang. Variabel pertama yang akan dilakukan analisis adalah ‘Year’. Kami melakukan analisis terhadap tahun dengan menghitung rata-rata tahun rilis dari setiap film pada masing-masing SVOD. Didapatkan bahwa rata-rata tahun rilis film untuk Netflix memiliki nilai yang paling tinggi, yaitu 2013.3, disusul dengan Hulu yang memiliki nilai rata-rata 2011.52, kemudian Prime Video dengan rata-rata 1999.51, dan rata-rata yang paling rendah adalah Disney+ sebesar 1997.79. Hal ini dapat diartikan bahwa rata-rata film yang ada di Netflix merupakan film yang dirilis sekitar tahun 2011 dan dibandingkan dengan SVOD yang lain, Netflix memiliki rata-rata terbesar untuk tahun rilis film.

Variabel kedua yang dilakukan analisis adalah umur untuk menentukan jumlah film berdasarkan batasan umur. Kami melihat value untuk data pada column umur terbagi menjadi '13+', '18+', '7+', nan, 'all', dan '16+'. Hal yang kami lakukan selanjutnya adalah membuat list yang digunakan untuk visualisasi terhadap setiap SVOD dan menghitung jumlah dari

masing-masing golongan umur dari Setiap SVOD. Didapatkan bahwa untuk golongan umur ‘all’ yang mencangkup semua umur, Disney+ dapat dikatakan memiliki jumlah film yang banyak, jika kita melihat jumlah film secara keseluruhan di Disney+ pada data ini tergolong sedikit. Untuk golongan umur lainnya, '13+', '18+', '7+', dan '16+, Prime Video tetap memiliki jumlah film yang paling banyak, dikarenakan dari data secara keseluruhan Prime Video secara jumlah data mendominasi dibandingkan dengan data yang SVOD yang lain.

Variabel selanjutnya adalah rating film yang ada di platform tersebut. Rating film yang kami ambil adalah Rating IMDb karena Rating dari Rotten Tomatoes terlalu banyak memiliki *missing values*. Kami menjumlahkan semua rating IMDb per platform lalu membaginya dengan jumlah film yang memiliki data rating IMDb. Maka, kami mendapatkan hasil bahwa Netflix memiliki rata-rata 5.82, Hulu memiliki rata-rata 5.69, Prime Video memiliki rata-rata 5.31, dan Disney+ memiliki rata-rata 5.99. Berdasarkan rating IMDb, Disney memegang rata-rata tertinggi.

Variabel selanjutnya adalah genre film yang dimiliki 4 platform tersebut. Pada data yang kami memiliki, beberapa film memiliki lebih dari satu genre. Maka, kami melakukan pemisahan genre terlebih dahulu agar mempermudah pengelompokan film berdasarkan genre. Lalu, kami hitung berapa film yang ada dalam 1 genre tersebut. Setelah itu, dapat terliahat bahwa Netflix memiliki genre utama, yaitu Drama, Comedy, Thriller, Romance, Action. Hulu memiliki genre Drama, Comedy, Thriller, Romance, Documentary. Prime Video memiliki genre Drama, Comedy, Thriller, Action, Documentary. Sedangkan, Disney memiliki warna genre yang lebih berbeda dari platform lainnya, yaitu Family, Comedy, Adventure, Fantasy, Drama.

Variabel yang akan dilakukan analisis selanjutnya adalah negara atau pada column data dituliskan ‘Country’. Seperti halnya data pada genre film, data negara juga memiliki lebih dari satu negara untuk satu judul film. Dengan metode yang sama, yaitu dengan melakukan pemisahan negara terlebih dahulu untuk memudahkan pengelompokan film berdasarkan negara, analisis dari setiap SVOD terhadap film dari suatu negara dapat dilakukan. Setelah melakukan ploting untuk mempermudah dalam melihat top negara dari setiap SVOD, didapatkan bahwa untuk Netflix, negara asal dari film yang dirilis pada SVOD tersebut adalah United States, India, United Kingdom, Canada, France. Untuk SVOD Hulu, negara yang paling mendominasi untuk film yang dirilis adalah United States, United Kingdom, Canada, France, Germany. Sementara untuk Prime Video, United States, United Kingdom, Canada, India, France menjadi negara asal dari film yang paling banyak d rilis pada SVOD tersebut. Film yang dirilis di Disney+ didominasi oleh film dari negara asal United States, United Kingdom, Canada, Australia, France.

Variabel terakhir yang akan dilakukan analisis adalah bahasa. Pada column ‘Language’, seperti halnya data pada genre film dan juga negara, data bahasa juga kembali memiliki lebih dari satu bahasa untuk satu judul film. Dengan metode pemisahan dan pengelompokan yang sama dengan data untuk genre dan negara, analisis dari setiap SVOD terhadap film dari variabel bahasa dapat dilakukan. Setelah melakukan plotting dengan bar plot untuk mempermudah dalam melihat top bahasa dari setiap SVOD, untuk Netflix, film yang dirilis pada SVOD tersebut didominasi oleh bahasa English, Hindi, Spanish, French, German. Sementara Untuk SVOD Hulu, bahasa yang paling mendominasi untuk film yang dirilis adalah English, French, Spanish, German, Japanese. Untuk bahasa English, French, Spanish, Hindi, Italian, Prime Video banyak merilis film dengan bahasa tersebut. Film yang dirilis di Disney+ didominasi oleh film dari bahasa English, French, Spanish. Hindi, Italian.

PENUTUP

Dari hasil analisa kami, didapat kesimpulan:

- **Netflix** merupakan SVOD yang paling banyak dibicarakan oleh masyarakat dan memiliki rata-rata tahun rilis film terbaru.
- **Hulu** merupakan SVOD yang memiliki film dengan tahun rilis terbaru setelah netflix
- **Prime Video** merupakan SVOD dengan jumlah film paling banyak namun rata-rata rating IMDbnya rendah
- **Disney+** merupakan SVOD yang berbasis *family friendly* dengan rata-rata tahun rilis film terlama serta rata-rata rating IMDb tertinggi namun jumlah film yang ada masih paling sedikit di antara platform SVOD lainnya.

Pemilihan SVOD yang ingin dipakai dapat didasarkan kebutuhan dan preferensi masing-masing pengguna.